2-25-2016

Bacteriophage EMS9: Preliminary Genomic Description

Hallie Rae Zimmer
Bowling Green State University, zimmerh@bgsu.edu

Bacteriophage EMS9: Preliminary Genomic Description

Hallie Zimmer

Abstract

EMS9 is a bacteriophage that was recently isolated from an *Escherichia coli* strain present in horse feces. Bacteriophage EMS9 consists of 98,771 base pairs that are organized into 139 predicted open reading frames (ORFs). These predicted genes potentially encode specific bacteriophage proteins. The genomic sequence of bacteriophage EMS9 is arranged intro three groups: early, middle, and late genes. Considerable homology between the ORFs of bacteriophage EMS9 and bacteriophages T5 and H8 exists. All of these bacteriophages are believed to use a rare two-step transfer mechanism to invade host cells. This annotation of the genomic sequence of EMS9 will provide a foundation for a further gene-by-gene analysis and comparison to other similar bacteriophages.

Introduction

Bacteriophages are viruses that invade and reproduce inside of bacterial cells (Summers, 2001). Bacteriophages greatly surpass bacteria in number and are the most numerous microorganisms in the world. Due to their highly adaptive nature, bacteriophages are also potentially the most diversified microorganisms (Labrie et al., 2010). The genetic material of bacteriophages tends to be shorter and less complex than that of bacteria or eukaryotes; therefore, the first genomic sequences to be determined belonged to a bacteriophage. Individual bacteriophage usually share some genetic similarities with other bacteriophages due to their adaptive nature and horizontal gene transfer. However, bacteriophage genomes also usually contain multiple novel genetic regions that have no known function (Hatfull, 2008).

One of the first bacteriophages studied is T5. Comparison of the complete genomic sequence of T5 (Wang et al., 2005) with the more recently identified bacteriophage H8 (Wolfgang et al., 2007) indicates that these two viruses are closely related. Phenotypically, they share an interesting mechanism for injecting their DNA and gaining access to host cells. Some of the strongest known promoters are present in bacteriophage T5 (Gentz and Bujard, 1985: Von Gabain and Bujard, 1975). These strong promoters lead to the production of pre-early genes that have detrimental effects on the host cell (Lanni and McCorquodale, 1968: Zweig et al., 1972). Other sets of promoters are subsequently engaged in a sequential process to complete various steps in the T5 life cycle (Sayers, 2006). An interesting and somewhat unique two-step transfer mechanism is used by bacteriophage T5 to invade host cells (Wang et al., 2005). 120 out of the 143 open reading frames of bacteriophage H8 are homologous to the open reading frames of bacteriophage T5 (Wolfgang et al., 2007). A number of these homologous genes encode products for which a function has yet to be determined.

The topic of this project is bacteriophage EMS9, recently isolated by our lab from an *E. coli* strain present in horse feces (Beck and Larsen, unpublished). The genome of EMS9 has been solved (Szuter and Larsen, unpublished) and is found to bear strong similarity with both bacteriophage T5 and bacteriophage H8. Some of the open reading frames for bacteriophage EMS9 are homologous to bacteriophage T5, bacteriophage H8, or both. This project focuses on the annotation of the EMS9 genome.

Gene sequences are the form of communication for all living things, with bacteriophage providing a simple, short story (relative to the much larger genomes of bacteria and higher organisms). The successful interpretation of this genetic language can allow for the further understanding of bacteriophage EMS9, bacteriophages similar to bacteriophage EMS9 that have already been sequenced, and bacteriophages similar to bacteriophage EMS9 that have yet to be identified or sequenced. A better understanding of bacteriophages on a genetic level can lead to the discovery of the function behind genomic regions of unknown function among many bacteriophages.

Results

One hundred thirty-nine putative ORFs were identified for bacteriophage EMS9. Figure 1 shows a genomic map of bacteriophage EMS9. The ORFs are summarized in Table 1. Here, the number of amino acids predicted to comprise the protein encoded by each numbered ORF is listed first, followed by the position (in base pairs) where the ORF occurs, the DNA strand on which the information is encoded and finally, the nature of the predicted protein is listed. Certain of the ORFs listed encode proteins that are predicted to function in a variety of processes important for reproduction of the virus, such as DNA replication, repair, lysis, or metabolism. Other ORFs encode structural proteins that will comprise the new viruses to be formed. Finally, some ORFs encode proteins involved in the maturation and release of new viruses from the host cell. While many of these ORFs contain information to encode proteins similar to those in the well-studied T5 bacteriophage and/or its relative H8, thirteen of the 139 ORFs encode hypothetical proteins that have yet to be identified in any other bacteriophage or living thing.

The genome of bacteriophage EMS9 is divided into early, middle, and late genes. A large amount of the early genes in bacteriophage EMS9 are homologous to bacteriophage T5. The early genes in bacteriophage T5 are transferred into a host cell along with genes that are probable inhibitors that disable the functions of the host cell (Wang et al., 2005). Middle genes tend to function in DNA replication, repair, lysis, or metabolism. For example, replication origin binding protein, replicative DNA helicase, DNA replication primase, DNA polymerase, and recombination endonucleases are some of the middle genes involved in DNA replication, repair, or lysis. Middle genes, such as HNH endonuclease, can also serve regulatory functions or have a role in signal transduction. In addition, NAD-independent DNA ligase subunits A and B are also found among the middle genes. NAD-independent DNA ligase is often present in most prokaryotic organisms as a single coding sequences (Wang et al., 2005); however, this gene has been divided into two subunits in the genome of bacteriophage EMS9.

Late genes mainly serve structural purposes for the bacteriophage. Structural genes are involved in phage infection, development, and protection from host disruption (Wang et al., 2005). Various tail proteins, including tail fiber protein, pore-forming tail protein, and receptor-binding tail protein, as well as various head proteins are found in the late genes of bacteriophage EMS9. It is also interesting to note that the ORFs that encode the A1 and A2 proteins that are found in the early genes of bacteriophage T5 are found near the end of bacteriophage EMS9 (ORFs 134 and 136). The reason for this location is unclear but is similar for the bacteriophage H8. The A1 and A2 early proteins of bacteriophage T5 are required for the second step of the two-step transfer mechanism of invading host cells to occur (Snyder and Benzinger, 1981). The A1 protein encodes for the completion of DNA transfer into the host cell, termination of the expression of early genes, and the degradation of host DNA. The A2 protein also encodes for the

completion of DNA transfer into the host cell as well as binds to host DNA, RNA polymerase, and lipopolysaccharide (Wang et al., 2005). Because A1 and A2 are essential for the initial transfer of DNA in T5, one possible explanation for the position of the ORFs encoding these proteins in EMS9 and H8 is that the DNA is inserted from the opposite end as occurs with T5.

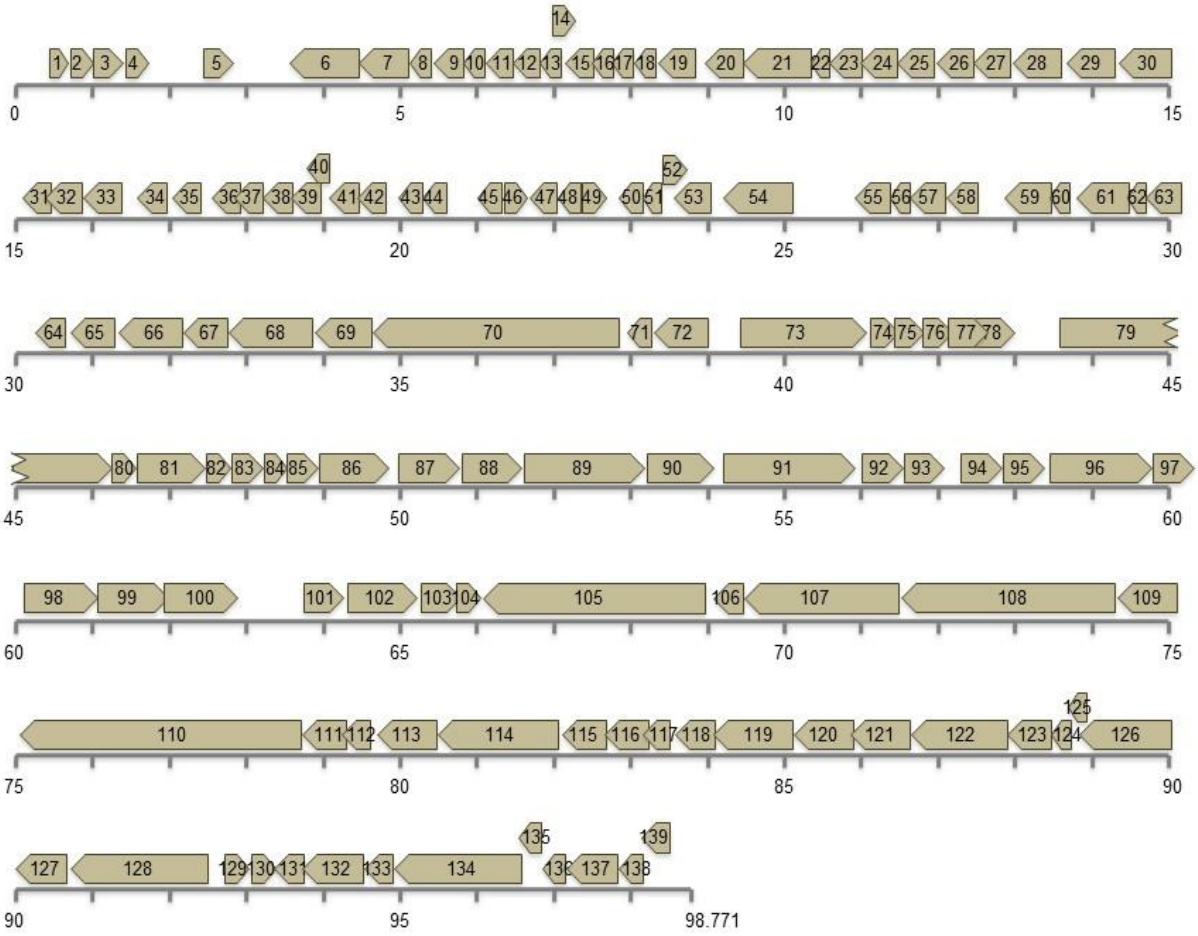

Figure 1. Genomic map of bacteriophage EMS9. The transcriptional direction of genes is indicated by the direction of the arrows.

Table 1
Characterization of putative EMS9 genes. Genes are listed sequentially by ORF number. The length of the predicted protein product is listed under "aa seq" as the number of amino acids (aa) that would comprise that protein. The precise location of the gene is listed under "Coordinates" as the base pair numbers that mark the ends of the ORF, with the DNA strand that contains the coding information identified as positive or negative under the heading of "Strand." The nature of the predicted protein is indicated under "Description," with a prefix of "T5" indicating that the predicted gene product is homologous to a particular protein (designated by number) encoded by T5.

Gene	aa seq	Coordinates	Strand	Description
1	59 aa	536-715	positive	T5 .013
2	70 aa	718-930	positive	hypothetical protein
3	113 aa	1054-1395	positive	T5 .015
4	64 aa	1499-1693	positive	T5 .016
5	103 aa	2561-2872	positive	hypothetical protein
6	294 aa	3648-4532	negative	T5 .018
7	201 aa	4610-5215	negative	T5 .019
8	61 aa	5215-5400	negative	hypothetical protein
9	160 aa	5400-5882	negative	T5 orf 22
10	67 aa	5842-6045	negative	T5 orf 23
11	123 aa	6102-6473	negative	T5p024
12	153 aa	6421-6882	negative	T5p025
13	69 aa	6879-7079	negative	T5 .027
14	69 aa	6995-7204	positive	T5.026
15	108 aa	7179-7505	negative	T5.028
16	81 aa	7495-7740	negative	T5 .029
17	96 aa	7737-7919	negative	T5 .030
18	89 aa	8041-8310	negative	T5p032/2C from T5
19	143 aa	8371-8802	negative	T5 .033
20	173 aa	8970-9491	negative	T5 .034
21	287 aa	9491-10354	negative	T5 .035

22	81 aa	10357-10602	negative	T5 .036
23	105 aa	10700-11017	negative	Thioredoxin
24	168 aa	10983-11489	negative	HNH Endonuclease
25	142 aa	11486-11914	negative	T5p027
26	137 aa	11992-12405	negative	T5 .039
27	137 aa	12483-12896	negative	Lysozyme
28	227 aa	12893-13576	negative	Holin/Lysis protein
29	199 aa	13707-14306	negative	Clp protease
30	250 aa	14319-15071	negative	deoxynucleoside-5'-monophosphate
31	117 aa	15071-15424	negative	hypothetical protein
32	147 aa	15355-15798	negative	T5 orf 041
33	226 aa	15795-16475	negative	T5 orf 043
34	114 aa	16629-17009	negative	T5p047
35	94 aa	17084-17368	negative	T5p048
36	115 aa	17605-17952	negative	T5p051
37	93 aa	17821-18102	negative	T5p052
38	115 aa	18179-18526	negative	T5p053
39	104 aa	18648-18962	negative	T5 orf 053
40	53 aa	18883-19044	negative	enterophage DT57C
41	122 aa	19044-19412	negative	acetyltransferase-related protein
42	96 aa	19520-19810	negative	T5p056
43	69 aa	20085-20294	negative	T5 .062
44	75 aa	20297-20524	negative	Cor
45	66 aa	21073-21273	negative	hypothetical protein
46	76 aa	21299-21529	positive	hypothetical protein
47	115 aa	21847-22194	negative	T5 .067
48	61 aa	22303-22488	negative	T5 .068

49	57 aa	22509-22682	positive	AGC_0074
50	68 aa	22930-23136	negative	T5 .073
51	55 aa	23129-23296	negative	T5 .074
52	68 aa	23285-23491	positive	hypothetical protein
53	189 aa	23503-24072	negative	T5 .076
54	315 aa	24235-25182	negative	T5 .080
55	152 aa	25954-26412	negative	T5 .081
56	81 aa	26401-26646	negative	T5 .082
57	149 aa	26651-27100	negative	T5 .083
58	105 aa	27106-27423	negative	T5 orf 079
59	213 aa	27866-28507	negative	T5 .085
60	60 aa	28557-28739	negative	T5p084
61	233 aa	28811-29512	negative	T5 orf 082
62	295 aa	29541-29777	negative	T5p086
63	171 aa	29820-30335	negative	T5p087
64	92 aa	30419-30697	negative	T5 .090
65	158 aa	30774-31250	negative	ribonuclease H
66	279 aa	31382-32221	negative	Thymidylate synthase
67	173 aa	32221-32742	negative	Dihydrofolate reductase
68	381 aa	32751-33896	negative	ribonucleoside reductase
69	163 aa	33963-34454	negative	HNH Endonuclease
70	1151 aa	34519-37974	negative	ribonucleoside reductase
71	65 aa	38014-38211	negative	hypothetical protein
72	250 aa	38213-38974	negative	phosphate starvation inducible protein
73	608 aa	39370-41196	positive	ribonucleoside reductase
74	70 aa	41296-41508	positive	hypothetical protein
75	76 aa	41505-41735	positive	T5p099

76	75 aa	41833-42060	positive	Sir2-like protein
77	142 aa	42063-42491	positive	T5 orf 098
78	130 aa	42501-42893	positive	T5p105
79	955 aa	43531-46398	positive	replication origin binding protein
80	80 aa	46373-46615	positive	T5p108
81	234 aa	46685-47389	positive	D2 protein
82	91 aa	47358-47633	positive	T5 .112
83	136 aa	47737-48147	positive	T5p111/D3 protein
84	95 aa	48164-48451	positive	T5 orf 105
85	102 aa	48502-48810	positive	T5 orf 106
86	323 aa	48898-49869	positive	NAD-dependent DNA ligase subunit A
87	259 aa	50072-50851	positive	NAD-dependent DNA ligase subunit B
88	255 aa	50844-51611	positive	D5 protein
89	507 aa	51643-53166	positive	replicative DNA helicase
90	296 aa	53163-54053	positive	DNA replication primase
91	570 aa	54206-55918	positive	DNA polymerase
92	138 aa	56172-56588	positive	hypothetical protein
93	131 aa	56717-56956	positive	hypothetical protein
94	130 aa	57399-57791	positive	DNA polymerase
95	165 aa	57784-58281	positive	T5p121
96	402 aa	58278-59486	positive	ATP dependent helicase/D10 protein
97	129 aa	59745-60134	positive	ATP dependent helicase/D10 protein
98	257 aa	60127-60900	positive	T5p124/D11 protein
99	325 aa	60937-61914	positive	recombination endonuclease
100	265 aa	61937-62733	positive	exonuclease subunit 2
101	160 aa	63737-64219	positive	D14 protein
102	291 aa	64219-65094	positive	flap endonuclease

103	148 aa	65091-65537	positive	deoxyUTP pyrophosphatase
104	69 aa	65554-65763	positive	T5p130
105	1004 aa	66028-69042	negative	tail fiber protein
106	140 aa	69042-69464	negative	T5 orf 140aa/phage tail protein
107	706 aa	69469-71589	negative	tail protein Pb4
108	949 aa	71589-74438	negative	tail protein Pb3
109	204 aa	74435-75049	negative	T5 .139/tail protein Pb9
110	1227 aa	75159-78842	negative	pore-forming tail protein Pb2
111	166 aa	78927-79427	negative	T5 .142
112	72 aa	79387-79605	negative	T5p139
113	299 aa	79758-80657	negative	tail protein gp24
114	464 aa	80662-85286	negative	tail protein gp25
115	161 aa	82083-82568	negative	T5p142
116	188 aa	82572-83138	negative	T5p143
117	68 aa	83132-83338	negative	T5p143
118	170 aa	83338-83850	negative	T5p144
119	458 aa	83910-85286	negative	major head protein Pb8
120	210 aa	85304-85936	negative	prohead protease
121	245 aa	85940-86677	negative	tail protein/head protein Pb10
122	405 aa	86674-87891	negative	portal protein
123	145 aa	87891-88328	negative	T5p149
124	56 aa	88394-88567	negative	hypothetical protein
125	47 aa	88683-88835	negative	T5.154
126	438 aa	88847-90163	negative	terminase large subunit
127	160 aa	90163-90645	negative	T5 .156
128	645 aa	90645-92582	negative	receptor-binding tail protein
129	42 aa	92807-92935	positive	T5p155

130	64 aa	93033-93227	positive	T5 .161
131	85 aa	93196-93453	negative	Escherichia phage EPS7 ACG0171
132	243 aa	93630-94361	negative	deoxynucleoside-5'-monophosphate
133	133 aa	94436-94837	negative	T5p160
134	554 aa	94907-96571	negative	A1
135	71 aa	96669-96914	negative	T5p163
136	138 aa	96924-97340	negative	A2
137	127 aa	97400-97783	negative	T5p165
138	56 aa	97929-98099	negative	hypothetical protein
139	99 aa	98077-98376	negative	T5.009

Discussion

Considerable homology of ORFs exists between bacteriophage EMS9 and bacteriophages T5 and H8. Many of the ORFs that these bacteriophages have in common are also located in the same genomic regions of each bacteriophage. The generation of such similar bacteriophages is likely caused by both homologous and non-homologous recombination (Hendrix, 2002). In addition, the existence of such similar bacteriophages supports the hypothesis that all bacteriophages evolved from the same genetic pool (Blaisdell et al., 1996; Hendrix, 2002).

The similarity of ORFs between bacteriophage EMS9 and bacteriophages T5 and H8 as well as the presence of proteins A1 and A2 in bacteriophage EMS9 strongly suggest that bacteriophage EMS9 may also share the interesting two-step transfer mechanism for invading host cells. The use of a two-step transfer mechanism by these bacteriophages advocates that a structure forms to prevent the left sequence and, therefore, the entire genome from being injected into the host cell at once. Repeat sequences that are capable of forming hairpin structures present in the early genes of the bacteriophage can function in the formation of the stop structure (Wang et al., 2005). Palindromic repeats as well as repeats in the injection-stop signal sequence may also facilitate the two-step transfer mechanism used by these bacteriophages (Heusterspreute et al., 1987).

In the two-step transfer mechanism, the remainder of the phage DNA enters the host cell and the middle genes are expressed only after the expression of the early genes. In the case of bacteriophage T5, the expression of middle genes occurs approximately 5 minutes after infection of the host cell, continuing for about an additional 20 minutes or until lysis occurs. The expression of late genes occurs around 10 to 12 minutes after infection and will continue until lysis occurs. The reason for this two-step DNA injection process is unclear, but it has been suggested that this allows the bacteriophage to evade defensive host processes that are destructive to bacteriophage replication (Wang et al., 2005).

Materials and Methods

This project considered a complete genome of the bacteriophage EMS9, originally isolated from horse feces and propagated on the K-12 *E. coli* strain W3110 (Beck and Larsen, unpublished). Purified bacteriophage DNA was then commercially analyzed by GeneWiz Corporation using an NGS ion-torrent sequenator. The resultant raw data was assembled to a solved sequence in the Larsen lab using a CLC genomic workbench software package (Szuter and Larsen, unpublished). I identified potential open reading frames using the online freeware "NEBcutter," DNA Master software from Dr. J.G. Lawrence at the University of Pittsburgh, and the CLC Main Workbench software package (Qiagen Corp.). The predicted proteins from each ORF were screened against the protein database at the National Center for BioInformatics (NCBI) using the online protein BLAST function provided by NCBI to identify similar proteins.

Prospectus

The completion of the preliminary genomic description of bacteriophage EMS9 will allow for the further in depth gene-by-gene analysis of bacteriophage EMS9. The preliminary genomic description strongly suggests that bacteriophage EMS9 follows the same genomic model as bacteriophages T5 and H8. The similarity between these bacteriophages as well as their interesting and unique two-step transfer mechanism makes bacteriophage EMS9 worthy of further inspection and analysis. A further genetic analysis of bacteriophage EMS9 and the comparison to other similar bacteriophages can allow for a deeper understanding of the evolutionary strategies of bacteriophages.

References

Blaisdell, B.E., Campbell, A.M. and Karlin, S. “Similarities and dissimilarities of phage genomes.” *PNAS* 93 (1996): 5854-5859.

Gentz, R. and Bujard, H. “Promoters recognized by *Escherichia coli* RNA polymerase selected by function: highly efficient promoters from bacteriophage T5.” *Bacteriol* 164 (1985):70-77.

Hatfull, Graham F. "Bacteriophage Genomics." *Current Opinion in Microbiology* 11.5 (2008): 447-53.

Hendrix, R.W. “Bacteriophages: evolution of the majority.” *Theoretical Population Biology* 61 (2002): 471-480.

Heusterspreute, M., Ha-Thi, V., Tournis-Gamble, S. and Davison, J. “The first-step transfer-DNA injection-stop signal of bacteriophage T5.” *Gene* 52 (1987): 155-164.

Labrie, S.J., Samson, J.E., and Moineau, S. "Bacteriophage Resistance Mechanisms." *Nature Reviews Microbiology* 8.5 (2010): 317-27.

Lanni, Y.T. and McCorquodale, D.J. “DNA metabolism in T5-infected *Escherichia coli*: biochemical function of a presumptive genetic fragment of the phage.” *Virology* 19 (1968): 72.

Sayers, J.R. “Bacteriophage T5” in The Bacteriophages (R. Calendar, ed), Oxford University Press (Oxford, Great Britain), (2006), pp268-276.

Snyder, C.E. and Benzinger, R.H. "Second-Step Transfer of Bacteriophage T5 DNA: Purification and Characterization of the T5 Gene A2 Protein." *Journal of Virology* 40.1 (1981): 248-57.

Summers, William C. "Bacteriophage Therapy." *Annual Reviews in Microbiology* 55.1 (2001): 437-51.

Von Gabain, A. and Bujard, H. “Interaction of *Escherichia coli* RNA olymerase with promoters of several coliphage and plasmid DNAs.” *PNAS* 76 (1975):189-193.

Wang, J. et al. "Complete Genome Sequence of Bacteriophage T5." *Virology* 332.1 (2005): 45-65.

Wolfgang, R. et al. "FepA- and TonB-Dependent Bacteriophage H8: Receptor Binding and Genomic Sequence." *Journal of Bacteriology* 189.15 (2007): 5658-74.

Zweig, M., Rosenkranz, H.S., and Morgan, C. “Development of coliphage T5: ultrastructural and biochemical studies.” *Virology* 9 (1972): 526-543.